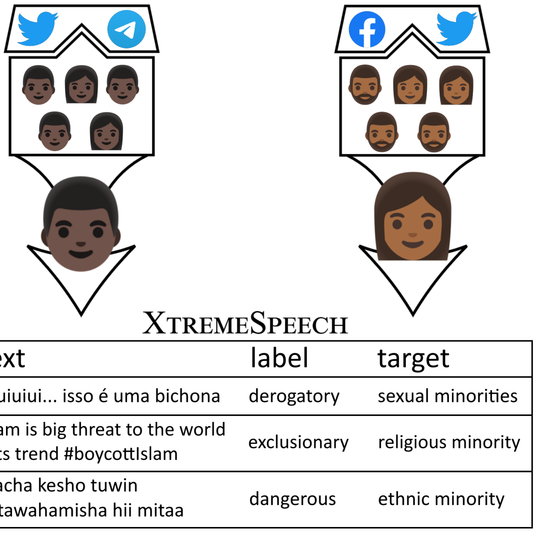
Listening to Affected Communities to Define Extreme Speech: Dataset and Experiments

Antonis Maronikolakis[1*] **Axel Wisiorek**[1,2] **Leah Nann**[3] **Haris Jabbar**[1]
Sahana Udupa[3] **Hinrich Schütze**[1]

[1]CIS, Center for Information and Language Processing
[2]Center for Digital Humanities [3]Institute of Social and Cultural Anthropology
LMU Munich
[*]*antmarakis@cis.lmu.de*

Abstract

Building on current work on multilingual hate speech (e.g., Ousidhoum et al. (2019)) and hate speech reduction (e.g., Sap et al. (2020)), we present XtremeSpeech,[1] a new hate speech dataset containing 20,297 social media passages from Brazil, Germany, India and Kenya. The key novelty is that we directly involve the affected communities in collecting and annotating the data – as opposed to giving companies and governments control over defining and combatting hate speech. This inclusive approach results in datasets more representative of actually occurring online speech and is likely to facilitate the removal of the social media content that marginalized communities view as causing the most harm. Based on XtremeSpeech, we establish novel tasks with accompanying baselines, provide evidence that cross-country training is generally not feasible due to cultural differences between countries and perform an interpretability analysis of BERT's predictions.



XtremeSpeech

text	label	target
Uluiuiui... isso é uma bichona	derogatory	sexual minorities
Islam is big threat to the world Lets trend #boycottIslam	exclusionary	religious minority
wacha kesho tuwin tutawahamisha hii mitaa	dangerous	ethnic minority

Figure 1: High-level overview of hate speech data collection. Instead of querying for data on our own, we work with fact-checkers advocating for targeted communities who collect and label data as they organically come across it online. We believe this inclusive approach results in datasets more representative of online speech marginalized communities are exposed to.

1 Introduction

Much effort has been put into curating data in the area of hate speech, from foundational work (Waseem and Hovy, 2016; Davidson et al., 2017) to more recent, broader (Sap et al., 2020) as well as multilingual (Ousidhoum et al., 2019) approaches. However, the demographics of those targeted by hate speech and those creating datasets are often quite different. For example, in Founta et al. (2018), 66% of annotators are male and in Sap et al. (2020), 82% are white. This may lead to unwanted bias (e.g., disproportionately labeling African American English as hateful (Sap et al., 2019; Davidson et al., 2019a)) and to collection of data that is not representative of the comments directed at target groups; e.g., a white person may not see and not have access to hate speech targeting a particular racial group.

An example from our dataset is the Kenyan social media post "… We were taught that such horrible things can only be found in Luo Nyanza." The Luo are an ethnic group in Kenya; Nyanza is a Kenyan province. The post is incendiary because it suggests that the Luo are responsible for horrible things, insinuating that retaliation against them may be justified. Only a group of people deeply rooted in Kenya can collect such examples and understand their significance.

XtremeSpeech. In this paper, we present XtremeSpeech, a new hate speech dataset containing 20,297 social media passages from Brazil, Germany, India and Kenya. The key novelty is that we empower the local affected communities (as opposed to companies and governments) to collect

[1] Code and data available at https://github.com/antmarakis/xtremespeech

and annotate the data, thus avoiding the problems inherent in approaches that hire outside groups for hate speech dataset creation. In more detail, we built a team of annotators from fact-checking groups from the four different countries. These annotators both collected and annotated data from channels most appropriate for their respective communities. They were also involved in all phases of the creation of XtremeSpeech, from designing the annotation scheme to labeling. Our inclusive approach results in a dataset that better represents content targeting these communities and that minimizes bias against them because fact-checkers are trained to be objective and know the local context. Figure 1 gives a high-level overview of data collection and annotation for XtremeSpeech.

XtremeSpeech also is a valuable resource because existing hate speech resources are not representative for problematic speech on a worldwide scale: they mainly cover Western democracies. In contrast, our selection is more balanced, containing three countries from the Global South and one Western democracy.

We present a data statement (see Bender and Friedman (2018)) in Appendix A.

Anthropological perspective. It has been argued that the NLP community does not sufficiently engage in interdisciplinary work with other fields that address important aspects of hate speech (Jo and Gebru, 2020). In this work, we take an anthropological perspective: the research we present is a collaboration of anthropologists and computational linguists. As a discipline that engages in the study of society and culture by exploring the lived worlds of people, and with a commitment to the application of knowledge to address human problems, sociocultural anthropology can provide a highlevel framework for investigating and theorizing about the phenomenon of hate speech and its cultural variations.

We also take an anthropological perspective for defining the terminology in this paper. Potentially harmful online speech is most often referred to by NLP researchers and general media[2] as **hate speech**. From its original, culturally-grounded meaning, *hate speech* has evolved into a primarily legal and political term with different definitions, depending on who uses it (Bleich, 2014; Saltman and Russell, 2014; Bakalis, 2018). We therefore use the concept of **extreme speech** from anthropology and adopt its definition as *speech that pushes the boundaries of civil language* (Udupa and Pohjonen, 2019; Udupa et al., 2021). In investigating extreme speech, anthropologists focus on cultural variation and historical conditions that shape harmful speech.

Extreme speech categories. We differentiate between extreme speech that requires removal (**denoted R**) and speech for which moderation (**denoted M**) is sufficient. Extreme speech of the M category consists of **derogatory** speech – roughly, disrespectful and negative comments about a group that are unlikely to directly translate into specific harm. We further subdivide R extreme speech into **exclusionary** extreme speech (roughly: speech inciting discrimination) and **dangerous** extreme speech (roughly: speech inciting violence); definitions are given in §3.2. This distinction is important when considering removal of extreme speech; e.g., dangerous speech may warrant more immediate and drastic action than exclusionary speech.

XtremeSpeech does not contain neutral text, focusing solely on M and R extreme speech. Neutral text has been shown to be easier to label both for humans and models while identifying and subclassifying non-neutral text (i.e., extreme speech) remains the Achilles' heel of NLP models (Davidson et al., 2017; Ranasinghe and Zampieri, 2020).

Finally, we also annotate the targets of extreme speech; examples are "religious minorities" and "immigrants" (frequent targets in India and Germany respectively).

Classification tasks. We define three classification tasks. (i) **REMOVAL.** The two-way classification M vs. R. (ii) **EXTREMITY.** The three-way classification according to degree of extremity: derogatory vs. exclusionary vs. dangerous. (iii) **TARGET.** Target group classification.

We propose a series of baselines and show that model performance is mediocre for REMOVAL, poor for EXTREMITY and good for TARGET. Further, we show that BERT-based models are unable to generalize in cross-country and cross-lingual settings, confirming the intuition that cultural and world knowledge is needed for this task. Further, we perform a model interpretability analysis with LIME (Ribeiro et al., 2016) to uncover potential model biases and deficiencies.

Contributions. In summary, we **(i)** establish

[2] https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/

a community-first framework of data curation, **(ii)** present XtremeSpeech, a dataset of 20,297 *extreme speech* passages from Brazil, Germany, India and Kenya, capturing target groups and multiple levels of extremity, **(iii)** propose a series of tasks and baselines, as the basis for meaningful comparison with future work, **(iv)** show performance both for models and humans is low across tasks except in target group classification, **(v)** confirm the intuition that extreme speech is dependent on social and cultural knowledge, with low cross-lingual and cross-country performance.

2 Related Work

Earlier work in hate speech detection focused on data collection, curation and annotation frameworks (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018). Recent work has expanded the set of captured labels to include more pertinent information such as targets and other forms of abuse (Sap et al., 2020; Hede et al., 2021; Guest et al., 2021; Grimminger and Klinger, 2021; Ross et al., 2017) as well as benchmarking (Röttger et al., 2021; Mathew et al., 2021). Analysis of datasets has been performed too (Madukwe et al., 2020; Kim et al., 2020; Wiegand et al., 2019; Swamy et al., 2019; Davidson et al., 2019a).

Work has also been conducted to expand research to multiple languages (Ousidhoum et al., 2019; Ranasinghe and Zampieri, 2020; Ross et al., 2017; Nozza, 2021; Zoph et al., 2016; Marivate et al., 2020; Nekoto et al., 2020). XtremeSpeech contributes to this goal by providing Brazilian Portuguese, German, Hindi and Swahili data.

Research has also been conducted to investigate annotation bias and annotator pools (Al Kuwatly et al., 2020; Waseem, 2016; Ross et al., 2017; Shmueli et al., 2021; Posch et al., 2018), as well as bias (especially racial) in existing datasets (Davidson et al., 2019b; Laugier et al., 2021). It was found that data can reflect and propagate annotator bias. To address this, we diversify the annotator pool in our work.

In another line of work, theoretical foundations are being established, in the form of taxonomies (Banko et al., 2020), definitions (Wiegand et al., 2021; Waseem et al., 2017) and theory (Price et al., 2020; Laaksonen et al., 2020). We are adding to this with definitions based on fieldwork and grounded research, inspired by anthropological and ethnographic work that investigates the societal impact of online hate and extreme speech (Boromisza-Habashi, 2013; Donovan and danah boyd, 2021; Haynes, 2019; Udupa and Pohjonen, 2019; Hervik, 2019).

Further, strides have been made in the ethics of AI. Who should collect data and who is responsible for model deployment decisions? Calls have been made for more inclusive pools of annotators and domain experts overseeing NLP projects, as well as exploration of other ethical dilemmas (Leins et al. (2020); Jo and Gebru (2020); Mitchell et al. (2020); Vidgen et al. (2019); Gebru (2019); Mohamed et al. (2020), *inter alia*). With our focus on community-embedded fact-checkers our framework is more inclusive than previous work.

3 Dataset

3.1 Dataset Description

XtremeSpeech consists of 20,297 passages, each targeted at one or more groups (e.g., immigrants). Data is collected from Brazil, Germany, India and Kenya. Passages are written in Brazilian Portuguese, German, Hindi and Swahili, as well as in English. English can either be used on its own, or in conjunction with the local language in the form of code switching. We capture this in the annotation: passages that contain English –even if it is only a hashtag in a tweet– are marked as containing both languages. Table 1 shows the distribution of languages.

Further, XtremeSpeech is platform-agnostic, with text collected from multiple online platforms, as well as direct messaging (anonymized) from the third quarter of 2020 until the end of 2021. Namely, Brazilian annotators sourced WhatsApp groups, the German team collected data from Facebook, Instagram, Telegram, Twitter and YouTube, Indian annotators sourced Facebook and Twitter and the Kenyan annotators collected data from Facebook, Twitter and WhatsApp. While forms of extreme speech may originate from one place, dissemination to other platforms is swift (Rogers, 2020). Proprietary efforts have also taken a platform-agnostic approach.[3]

Passages were labeled both on content and target levels. On their content they are labeled as derogatory, exclusionary or dangerous. On the target level, we make a distinction between text targeted at protected groups and at institutions of

[3] https://www.perspectiveapi.com/

power. We take into account the following protected groups: ethnic minorities, immigrants, religious minorities, sexual minorities, women, racialized groups, historically oppressed caste groups, indigenous groups and large ethnic groups. We also give the annotators the option to input any other group. For institutions of power, possible targets are politicians, legacy media and the state. To allow for political discourse, extreme speech against institutions of power should not be filtered out, so such speech was marked as derogatory.

3.2 Extreme Speech Definitions

Building on Benesch (2018) and Udupa (2021), we define extreme speech labels as follows:[4]

Derogatory Extreme Speech: Text that crosses the boundaries of civility within specific contexts and targets either individuals/groups based on protected characteristics (eg., ethnicity and religious affiliation) or institutions of power (state, media, politicians). Includes derogatory expressions about abstract categories/concepts.

Exclusionary Extreme Speech: Text that calls for or implies exclusion of vulnerable groups based on protected attributes (for example, ethnicity, religion and gender). Exclusionary text marginalizes, delegitimizes and discriminates against target groups. Text targeting abstract concepts or institutions is not exclusionary, except when there is reason to believe that such attacks call for or imply the exclusion of vulnerable groups associated with these abstract concepts or institutions.

Dangerous Extreme Speech: Text that has a reasonable chance to trigger harm against target groups (eg., ostracism and deportation). All the following criteria should be met for a passage to be classified as dangerous: (i) content calls for harm, (ii) speaker has high degree of influence over audience, (iii) audience has grievances and fears that the speaker can cultivate, (iv) target groups are historically disadvantaged and vulnerable to harm, (v) influential means to disseminate speech.

Whereas derogatory extreme speech is a form of speech that *requires moderation but, generally, not removal* (denoted with M), exclusionary and dangerous speech are forms of speech that do *require removal* (denoted with R) in most cases to protect users from potential harm. We make a distinction between exclusionary and dangerous speech in order to introduce a more fine-grained scale of extremity that can dictate more focused policy (e.g., more severe punishment may be appropriate for dangerous speech). It has been shown in previous work that while neutral text is easier to detect (Davidson et al., 2017; Ranasinghe and Zampieri, 2020; Risch and Krestel, 2020), models find it hard to differentiate between different types of extreme speech (e.g., between our definitions of M or R, or between merely offensive versus hateful speech), a task challenging even for humans. By focusing on the difficult distinctions within non-neutral text, we hope to contribute to research that will be able to classify types of potentially harmful speech correctly in the future, which is both the critical point of extreme speech research and the main obstacle towards effective filtering.

Exemplary cases for the three labels (derogatory, exclusionary, dangerous) were discussed in detail with the annotators. We believe our interdisciplinary approach will lead to data more aligned with the real world and will benefit the target groups and communities to greater effect.

3.3 Data Collection

3.3.1 Annotator Profiles

We selected Brazil, Germany, India and Kenya to cover a range of cultures and communities. Each annotator is a fact-checker who i) is local, ii) is independent (i.e., not employed by social media companies or large media corporations) and iii) investigates the veracity of news articles, including articles directed at or related to local communities. There are 8 female and 5 male annotators (per country, female/male counts are 2/1 in Brazil, 4/0 in Germany, 2/2 in India and 0/2 in Kenya).

Fact-checking companies were scouted and individual fact-checkers interviewed by our anthropology team to verify their familiarity with extreme speech, their expertise in local community affairs and their ability to act as annotators in our project.

We see independent fact-checkers as a key stakeholder community that provides a feasible and meaningful gateway into cultural variation in online extreme speech. Through their job as fact-checkers, they regularly come in contact with extreme speech, with communities that peddle extreme speech as well as with communities targeted by extreme speech (further details in Appendix C).

3.3.2 Annotation Scheme

Through an online interface, data is entered as found in online media. This interface (in the form

[4] Definitions were shared as annotation instructions.

of a web page, see Appendix C.4) serves both as the data entry point and the annotation form. After finding a passage of extreme speech, annotators enter it in our form and are prompted to label it (see categories in §3.1).

3.4 Inter-annotator Agreement

To verify the quality of XtremeSpeech, we calculate inter-annotator agreement. The data collected from one annotator is shown to another for verification (details in Appendix C.2). Only the text passage is shown to annotators, without prior category assignments. The agreement scores we measure are: Cohen's kappa (κ, McHugh (2012)), Krippendorff's alpha (α, Krippendorff (2011)), intraclass correlation coefficient (two-way mixed, average score $\text{ICC}(3, k)$ for $k = 2$, Cicchetti (1994)) and accuracy (defined as the percentage of passages where both annotators agreed).

For the three extreme speech labels, $\kappa = 0.23$, $\alpha = 0.24$ and $\text{ICC}(3, k) = 0.41$ (considered "fair" (Cicchetti, 1994)). Accuracy is 63% overall, 78% for derogatory, 40% for exclusionary and 19% for dangerous. For the M vs. R task, accuracy is 78.4% for M and 46.3% for R. For the classification of the target of extreme speech, $\kappa = 0.69$.

Scores are low compared to other NLP tasks, which is unfortunately a widespread phenomenon in hate speech research. In Founta et al. (2018), only in 55.9% of passages did at least 4 out of 5 annotators agree. In Sap et al. (2020), the α score was 0.45, with a 76% agreement on "offensiveness" and 74% on "targeted group". In Davidson et al. (2017), there was a 90% agreement on whether text was neutral, offensive, or hateful. In Ross et al. (2017), a German dataset, α was between 0.18 and 0.29, while in Ousidhoum et al. (2019), a multilingual dataset, α was between 0.15 and 0.24.

We argue that in our work, not only are we dealing with a heavily imbalanced dataset, but also that the task is even more challenging than prior work, which collects both neutral passages and hate speech (e.g., in Davidson et al. (2017)). We only collect extreme speech, so whereas in prior work the annotators need to differentiate between neutral and extreme speech (a relatively easier task (Ranasinghe and Zampieri, 2020; Risch and Krestel, 2020)), our annotators only make decisions on the hard task of determining different degrees of extremity.

	Brazil	Germany	India	Kenya
Local	5109	4922	2778	405
English	0	6	1056	2695
Both	0	71	1174	2081

Table 1: Passages per country and language combination

3.5 Reannotation

After discussing inconsistently labeled passages with the annotators, we found that there was disagreement about groups currently in power, specifically, the Kikuyu and Kalenjin ethnic groups (more information in Appendix D). One annotator considered them ethnic minorities because most other ethnic groups are pitted against them. The other annotator did not view them as minorities because they are (i) the two most populous ethnic groups and (ii) are not in the minority when it comes to representation in positions of power. A consensus was reached to add a new target label, "large ethnic group", to correctly represent this state of affairs in the annotation.

As is common practice, instead of limiting the reannotation to passages the annotators disagreed on, we provided all potentially affected passages for reannotation, i.e., all "indigenous group" and "ethnic minority" passages.

3.6 Dataset Analysis

3.6.1 Extreme Speech Analysis

XtremeSpeech contains 20,297 passages from the four countries. From each country, we chose to only collect data on one local language plus English. The distribution of languages is shown in Table 1. While for Germany and Brazil, English is rarely used, in India and Kenya it is more prominent, both on its own and in code switching.

The distribution of labels, shown in Table 2, varies a lot from country to country. For example, in Germany annotators labeled far fewer passages as dangerous speech, reflecting stricter regulatory controls over speech compared to the other countries. Data is also heavily imbalanced in Brazil, with the majority of passages being derogatory.

The distribution of targets per country (shown in Table 4) again shows large divergences between countries. In Germany, immigrants are the main target group because of right-wing opposition to recent immigration. In India, religious minorities dominate the target group statistics because of the conflict between Hindus and Muslims. Thus

	Brazil	Germany	India	Kenya	*Total*
Der.	4774	2643	2225	3389	*13031*
Exc.	115	2340	1422	1024	*4901*
Dan.	220	16	1361	768	*2365*

Table 2: Distribution of extreme speech labels

XTREMESPEECH reflects a country's social and political situation to a reasonable extent.

3.6.2 Word Frequency

Table 3 shows the most frequent words for the three extreme speech labels for the four countries. We see that words indicative of sociopolitical conflict appear frequently: "comunista" and "feminista" in Brazil; "merkel" (a German politician) and "deutsche" (meaning: "German"), as well as the word for Jew, "jude" in Germany; words referring to religion (e.g., "muslims", "hindus") in India. In Kenya, political entities ("Ruto" and "Raila", names of two Kenyan politicians) as well as ethnic groups (e.g., "Kikuyus", "Kalenjins", two powerful groups in Kenya) are among the most frequent words, with ethnic groups appearing particularly prominently in the two forms of extreme speech that should be removed (R).

4 Experiments

We establish XTREMESPEECH baselines for large pretrained models and traditional machine learning models (details in Appendix E). As introduced in §1, we address three novel tasks: predicting the extremity of speech (EXTREMITY), whether a passage should be removed or not (REMOVAL) and the target of extreme speech (TARGET).

Unless noted otherwise, our measure is micro-averaged F1. We split each country set into train:dev:test at 80:10:10 rates, sampling equally for all labels.[5] In Tables 5, 6, 7, 8, 9 we show results on the development set (test set results are shown in Appendix G).

We evaluate both multilingual (mBERT, XLM-R (Conneau et al., 2020)) and monolingual (langBERT) models. Each monolingual model was pretrained on the local language we are using for each corresponding country; e.g., the Indian model was pretrained on Hindi. For finetuning and classification with BERT-based models, a task-specific head is added which takes as input the [CLS] token representation.

4.1 EXTREMITY Task

Table 5 shows that baseline performance is rather low in three-way classification (EXTREMITY). In India and Kenya performance is acceptable; in Germany as well if we exclude the dangerous label, which only has 16 passages. In Brazil, however, where the predominant class is derogatory speech (with more than 90% of all passages labeled as derogatory), performance is low, with no model managing to detect exclusionary speech.

XLM-R performs relatively poorly, only scoring competitively in the low-resource Kenyan set. langBERT is competitive for Brazil and Germany, less so for Kenya and performs badly for India. This can be explained by the divergence of pretraining and XTREMESPEECH text: all langBERT models are pretrained on a single language (Brazilian Portuguese, German, Hindi and Swahili respectively). In the Brazilian and German sets there is primarily only one language used so langBERT performs better in those sets, while it performs worse in countries where English is more predominant both as a standalone language and in code switching, which is the case for India and Kenya.

4.2 REMOVAL Task

Table 6 shows that results are overall better for the binary task M (moderation) vs. R (removal) than for the fine-grained EXTREMITY task. BERT-based models perform particularly well. mBERT performs especially well in the Indian set and the monolingual langBERT models again perform well in the Brazilian and German sets; this time we see improvements in the Kenyan set too. LSTMs perform well, in some instances competitively with transformers. XLM-R does not seem to compute good representations and performs poorly for all languages except for the low-resource Kenyan set.

4.3 TARGET Task

Table 7 shows that transformers are effective for the 8-way multilabel classification of target. In Table 3 and Table 10, we show top words according to frequency in the dataset and contribution to mBERT predictions in the EXTREMITY task respectively. Words denoting ethnicity ("kikuyu"), religion ("hindu", "Muslim") and gender ("puta", "girls") appear often and, not surprisingly, are reliable indicators of targeted groups, making this task easier than the other two.

[5] The German subset only contains 16 dangerous passages, so results for the particular label are of limited utility.

	Brazil	Germany	India	Kenya
Der.	puta, vai, filho, arrombada, pra, vc, comunista, cu, traveco, tomar	mehr, deutschland, merkel, schon, mal, ja, immer, deutsche, land, neger	के, नही, muslims, भीमटे, muslim, मुल्ल, hindu, india, देश, hindus	Ruto, people, Raila, know, ruto, Kenya, never, even, Uhuru, us
Exc.	puta, feminista, pra, bichona, ucranizar, nojenta, ser, marmita, bandido, cu	deutschland, mehr, darf, ja, antwort, land, deutschen, juden, deutsche, mal	muslims, hindu, देश, bhimte, india, भीम, hindus, भारत, मुल्ल, country	Kikuyus, Ruto, Kenya, kikuyu, Raila, people, never, Uhuru, Luos, Kalenjins
Dan.	fechar, stf, pra, povo, ucranizar, vai, q, ser, hora, bolsonaro	jude, europa, darf, juden, muslim, scheiss, freiheitskampf, völker, fällt, niemals	muslims, muslim, hindu, hindus, india, girls, love, देश, women, religion	Ruto, people, killed, Kikuyus, Raila, Kenya, know, Rift, must, time

Table 3: Most frequent words per label and country

	Brazil		Germany		India		Kenya		*Total*	
	n	%	n	%	n	%	n	%	*n*	*%*
Religious Minorities	16	0.5	1269	23.8	3522	64.7	111	2.2	*4918*	*25.4*
Any Other	1066	30.5	34	0.6	356	6.5	1534	30.3	*2990*	*15.5*
Immigrants	28	0.8	2355	44.1	109	2.0	292	5.8	*2784*	*14.3*
Women	1479	42.3	367	6.9	418	7.7	396	7.8	*2660*	*13.8*
Large Ethnic Groups	0	0.0	0	0.0	0	0.0	2273	44.8	*2273*	*11.8*
Sexual Minorities	674	19.3	347	6.5	89	1.6	80	1.6	*1190*	*6.2*
Historically Oppressed Caste Groups	45	1.3	1	0.0	853	15.7	33	0.7	*932*	*4.8*
Racialized Groups	78	2.2	527	9.8	3	0.1	80	1.6	*688*	*3.6*
Ethnic Minorities	58	1.7	430	8.1	89	1.6	77	1.5	*654*	*3.4*
Indigenous Groups	50	1.4	6	0.1	5	0.1	195	3.8	*256*	*1.3*

Table 4: Total number and percentage of messages directed at target groups

	Brazil			Germany			India			Kenya		
	Der.	Exc.	Dan.	Der.	Exc.	Dan.	Der.	Exc.	Dan.	Der.	Exc.	Dan.
Human	97.2	21.2	0.0	73.0	61.6	0.0	91.1	16.9	4.9	68.9	10.7	57.2
Majority	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
SVM	100.0	0.0	35.6	67.8	62.9	0.0	76.7	29.8	65.6	89.6	41.9	38.8
LSTM	98.4	0.8	0.0	59.4	68.6	0.0	56.3	64.8	0.0	64.9	63.4	0.0
langBERT	99.7	0.0	54.8	62.0	70.6	0.0	87.4	0.0	53.4	83.3	38.5	45.2
mBERT	98.9	0.0	49.3	56.3	72.4	0.0	60.9	45.5	81.3	83.5	48.4	48.8
XLM-R	100.0	0.0	0.0	58.7	76.4	0.0	89.1	6.7	56.1	88.3	46.9	40.0

Table 5: F1 on dev for EXTREMITY, the three-way extreme speech classification task

	Brazil		Germany		India		Kenya	
	M	R	M	R	M	R	M	R
Human	97.2	25.0	73.0	61.7	91.1	23.2	68.9	43.1
Majority	100.0	0.0	100.0	0.0	0.0	100.0	100.0	0.0
SVM	100.0	26.4	67.8	62.4	67.3	77.4	84.9	55.5
LSTM	98.4	20.8	57.8	71.5	61.9	80.2	86.1	46.8
langBERT	99.2	41.5	62.0	73.4	66.0	59.6	86.7	58.4
mBERT	100.0	30.3	61.1	69.1	66.7	78.8	81.7	61.9
XLM-R	100.0	0.0	100.0	0.0	0.0	100.0	82.0	61.9

Table 6: F1 on dev for REMOVAL, the two-way extreme speech classification task

	Brazil	Germany	India	Kenya
langBERT	95.4	92.1	85.5	83.1
mBERT	94.1	90.3	92.8	85.6
XLM-R	94.1	88.2	93.0	84.8

Table 7: LRAP (Label Ranking Average Precision) on dev for TARGET, the target group classification task

		Brazil			Germany			India			Kenya		
		Der.	Exc.	Dan.	Der.	Exc.	Dan.	Der.	Exc.	Dan.	Der.	Exc.	Dan.
train	Brazil	98.9	0.0	49.3	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Germany	94.1	0.0	0.0	56.3	72.4	0.0	80.0	30.8	0.0	82.9	29.0	0.0
	India	95.5	0.0	11.0	96.3	0.0	0.0	60.9	45.5	81.3	70.4	40.8	6.3
	Kenya	94.9	3.0	9.6	79.6	10.4	0.0	83.7	14.4	29.0	83.5	48.4	48.8

Table 8: F1 on dev for EXTREMITY in cross-country transfer (all languages)

		IN_{en}			KE_{en}		
		Der.	Exc.	Dan.	Der.	Exc.	Dan.
train	IN_{en}	60.0	44.8	0.0	60.9	50.8	0.0
	KE_{en}	85.0	0.0	18.8	78.2	61.9	74.5

Table 9: F1 on dev for EXTREMITY for cross-country transfer in English (IN/KE = India/Kenya)

4.4 Zero-Shot Cross-Country Classification

4.4.1 All languages

We evaluate mBERT on zero-shot cross-country transfer, i.e., training on one country and testing on the rest (results are shown in Table 8). Performance is in general poor, indicating that mBERT is not able to generalize from one country to another. Trained on Brazil, the model is unable to make any inferences on other countries. From Kenya to India, we see some transferability potential, with the model correctly identifying passages in all three classes (although at a non-competitively low rate). These results confirm our intuition that detecting extreme speech depends on social and cultural information, so zero-shot transfer, without access to specific information about the target country, is not a promising approach.

4.4.2 English

We investigate cross-country transfer of BERT, an English model. We only experiment with the two countries that have a nontrivial number of English passages, India (IN) and Kenya (KE), restricting the datasets to their English part only (denoted by IN_{en} and KE_{en} respectively). While cross-country performance is low for both countries, we see that $\text{KE}_{en} \rightarrow \text{KE}_{en}$ performance is high. We note that performance is better in $\text{KE}_{en} \rightarrow \text{KE}_{en}$ than in the previously examined $\text{KE}_{all} \rightarrow \text{KE}_{all}$ (where KE_{all} is the entire Kenyan set). This shows that for a single language within one country, BERT can indeed classify extreme speech with adequate accuracy.

4.5 Prediction analysis with LIME

To shed light on predictions of mBERT in the EXTREMITY task (described in §4.1) we extract top-contributing words with LIME (Ribeiro et al., 2016). Specifically, we compute the words that contribute the most to mBERT's predictions (alongside their weights) for each passage and then average the weights, returning the top 10 words with at least 5 occurrences in the examined set. This list is shown in Table 10.

Brazil	Germany	India	Kenya
fechar	Politiker	muslims	cows
Ucranizar	Grünen	Muslim	ruto
ucranizar	Mohammedaner	muslim	luo
safada	Juden	Muslims	wajinga
prender	Merkels	ko	kikuyu
lixo	Merkel	mullo	stupid
coisa	Regierung	Rohingyas	idiot
kkkkk	Opfer	ड	looting
Vagabundo	Islam	suvar	tangatanga
traveco	Moslems	डर	ujinga

Table 10: Top words contributing to predictions of mBERT for EXTREMITY

The Indian and German sets are dominated by religious groups ("Moslems", "Muslims"). In India, ethnic terms ("Rohingyas") are also present while in Germany we see extreme speech targeting politicians ("Merkel"). In Brazil we see politically divisive terms ("Ucranizar", a term originally meaning "Ukrainian Brazilian" which has now been appropriated to depict opponents to the right-wing as "communists") as well as insults like "traveco" (for "cross-dresser", used here as a slur). In Kenya, we see direct insults such as "idiot" and "wajinga" (meaning "foolish"), as well as ethnic group mentions such as "luo" and "kikuyu".

5 Conclusion

We have presented XTREMESPEECH, an extreme speech dataset, containing 20,297 passages from Brazil, Germany, India and Kenya. We capture both granular levels of extremity and targets of extreme speech by engaging a team of annotators from within the affected communities. In a collaboration of anthropologists and computational linguists, we established a community-based framework, with the goal of curating data more representative of real-world harms.

We introduce baselines for three novel tasks, including extreme speech and target group classification. We give experimental support for the intuition that extreme speech classification is dependent on cultural knowledge and that current NLP models do not capture this. Finally, we perform interpretability analysis on BERT's predictions to reveal potential deficiencies, showing that models rely heavily on keywords and names of marginalized groups.

We hope our community-driven work will contribute to the effective elimination of extreme speech against target groups, not just in Western democracies, but in a greater variety of countries worldwide.

6 Acknowledgments

This research has received funding from the European Research Council Proof of Concept grant (Agreement Number: 957442). For more about the project see https://www.ai4dignity.gwi.uni-muenchen.de.

The first author was partly supported by the European Research Council (#740516).

7 Ethical Considerations and Limitations

7.1 Ethics Statement

The data provided here contains extreme speech that can be shocking and harmful. We present this dataset as a way to peel back the veil of extreme speech against the selected under-represented communities around the world. We want to motivate the analysis of this overlooked area as a whole and the investigation of the various levels of extreme speech (derogatory, exclusionary and dangerous) as found in online social media. This data is not intended and should not be used for pretraining models applied to real-world tasks, since a model pretrained on this data could potentially exhibit and propagate the extreme speech found in the passages we collected.

Further, while we endeavored to include as many communities around the world as possible, the data we collected and the list of communities we included are of course non-exhaustive. For each country, we had a close circle of annotators, therefore it is possible other marginalized groups in these countries were not covered (although we made efforts to keep this to a plausible minimum).

7.2 Limitations

Due to limitations of both time and budget, we only gathered extreme speech without negative passages (ie. neutral language). These neutral passages form the majority of content on social media (Founta et al., 2018; Sap et al., 2020). Despite the abundance of such passages, annotating them using our current scheme would be time and effort-consuming (our annotators collect data on their own, from their own networks, without us querying and supplying data to them). Thus, to keep control in the hands of annotators while at the same time keeping their workload to a reasonable minimum, we decided to only collect extreme speech passages.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Chara Bakalis. 2018. Rethinking cyberhate laws. *Information & Communications Technology Law*, 27(1):86–110.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Susan Benesch. 2018. Dangerous speech: A practical guide.

Erik Bleich. 2014. Freedom of expression versus racist hate speech: Explaining differences between high court regulations in the usa and europe. *Journal of Ethnic and Migration Studies*, 40(2):283–300.

David Boromisza-Habashi. 2013. *Speaking Hatefully: Culture, Communication, and Political Action in Hungary*. Pennsylvania State University Press.

Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6:284–290.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019a. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019b. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*.

Bruno Heridet Delapouite. Accessed 10/11/2021. https://game-icons.net/.

Joan Donovan and danah boyd. 2021. Stop the presses? moving from strategic silence to strategic amplification in a networked media ecosystem. *American Behavioral Scientist*, 65(2):333–350.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press.

Timnit Gebru. 2019. Oxford handbook on ai ethics book chapter on race and gender.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online. Association for Computational Linguistics.

Nell Haynes. 2019. Writing on the walls: Discourses on bolivian immigrants in chilean meme humor. *International Journal of Communication*, 13(0).

Anushree Hede, Oshin Agarwal, Linda Lu, Diana C. Mutz, and Ani Nenkova. 2021. From toxicity in online comments to incivility in American news: Proceed with caution. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2620–2630, Online. Association for Computational Linguistics.

Peter Hervik. 2019. Ritualized opposition in danish practices of extremist language and thought. *International Journal of Communication*, 13(0).

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20, page 306–316, New York, NY, USA. Association for Computing Machinery.

Jae-Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *CoRR*, abs/2005.05921.

Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability.

Salla-Maaria Laaksonen, Jesse Haapoja, Teemu Kinnunen, Matti Nelimarkka, and Reeta Pöyhtäri. 2020. The datafication of hate: Expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data*, 3:3.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. Association for Computational Linguistics.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. In *Proceedings of the first workshop on Resources for African Indigenous Languages*, pages 15–20, Marseille, France. European Language Resources Association (ELRA).

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22:276–82.

Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 117–123, New York, NY, USA. Association for Computing Machinery.

Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy and Technology*, 33(4):659–684.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Lisa Posch, Arnim Bleier, Fabian Flöck, and Markus Strohmaier. 2018. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Julian Risch and Ralf Krestel. 2020. Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61, Marseille, France. European Language Resources Association (ELRA).

Richard Rogers. 2020. Deplatforming: Following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Erin Marie Saltman and Jonathan Russell. 2014. White paper – the role of prevent in countering online extremism.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Sahana Udupa. 2021. Digital technology and extreme speech: Approaches to counter online hate.

Sahana Udupa, Elonnai Hickok, Antonis Maronikolakis, Hinrich Schuetze, Laura Csuka, Axel Wisiorek, and Leah Nann. 2021. Artificial intelligence, extreme speech and the challenges of online content moderation.

Sahana Udupa and Matti Pohjonen. 2019. Extreme speech and global digital cultures. *International Journal of Communication*, 13(0).

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Data Statement

CURATION RATIONAL In our project, we venture to present a dataset on extreme speech across different countries (Brazil, Germany, India and Kenya). Fact-checkers from these countries were requested to gather and annotate data. These fact-checkers searched online platforms and communities to identify extreme speech based on their contextual language. The choice of sources was left to the fact-checkers, since they have intimate knowledge of the spread of extreme speech. Sources include social media (e.g., Twitter), fora (e.g., groups on Telegram) and direct messaging.

LANGUAGE VARIETY Data was collected for Brazilian Portuguese (pt-BR), German (de-DE), Hindi (hi-IN, either in the Devangari or Latin script), Swahili (sw-KE) and English used as a second language alongside these native languages.

SPEAKER DEMOGRAPHIC Speaker demographics were not recorded (and anonymized where necessary). Data was collected from Brazil, Germany, India and Kenya, so a fair assumption is that speakers come from these countries.

ANNOTATOR DEMOGRAPHIC Annotators were accredited fact-checkers in their respective countries. There were 8 female and 5 male annotators (per country, female/male counts are 2/1 in Brazil, 4/0 in Germany, 2/2 in India and 0/2 in Kenya). They were native speakers of (Brazilian) Portuguese, German, Hindi and Swahili. Ages were not recorded. Further (self-disclosed) information on annotators can be found at https://www.ai4dignity.gwi.uni-muenchen.de/partnering-fact-checkers/.

SPEECH SITUATION Speech consists entirely of text, posted and collected in 2020 and 2021. Text is mainly asynchronous, informal and spontaneous. Certain passages were posted as responses to other text (which was not collected) in a more synchronous manner. By the nature of this project, all passages contain a level of extremity.

TEXT CHARACTERISTICS Text comes from social media in the form of user comments. Length was limited to approximately two paragraphs (at the discretion of the annotators).

OTHER The team spanned multiple disciplines, ages and ethnicities.

	Brazil	Germany	India	Kenya	*Total*
Der.	15.8	22.5	26.0	24.2	*21.0*
Exc.	18.3	27.7	28.1	27.6	*27.6*
Dan.	21.2	40.5	30.3	29.6	*29.3*
Ovr.	16.1	25.0	27.8	25.7	*23.5*

Table 11: Average passage length statistics

B Data Analysis

B.1 Institutions of Power

Statistics of institutions of power are shown in Table 15. These groups can only be the target of derogatory speech, since we want to avoid censoring of speech aimed at these groups. Across all countries, we see that politicians are the predominant targets.

B.2 Average Passage Length

In Table 11 we show the average length of passages per label for each country. All sets show similar lengths, except Brazil where passages are overall shorter. Also, across sets, the more extreme a passage is, the longer it is on average.

C Annotation Details

C.1 Logistics

There are at least two annotators from each country. In some countries, we worked with fact-checker teams which themselves employ multiple fact-checkers. In these instances, annotation work was split according to the requirements and resources of the particular team. We ensured that all involved members were accredited fact-checkers and were interviewed by our anthropology team to verify they are familiar with extreme speech and are capable of identifying it. Payment was 1.5 Euros per passage provided for the original dataset and 1 Euro per passage for the re-annotation task.

C.2 Cross-annotation

In Table 12 we show the number of passages cross-annotated by each annotator. Annotators were split into two groups, A and B, according to availability and were tasked with cross-annotating the passages provided by the other group.

C.3 Inter-annotator agreement details

In Table 14 we show inter-annotator agreement scores per country. While Germany and Kenya have acceptable scores, the other two countries have low scores.

	Group A	Group B
Brazil	834 833	833
Germany	834 833	833
India	1250	417 417 416
Kenya	1250	1250

Table 12: Number of passages each group cross-annotated, evenly split among its members.

C.4 Online Interface

In Figure 2 we see the interface annotators used to enter and annotate data.

D Reannotation

After discussion with the annotators from Kenya, we found that there was disagreement surrounding two ethnic groups and the power dynamics around them. Namely, the Kikuyu and Kalenjin, two ethnic groups currently in power in Kenya. They make up around 17% (largest group) and 13% (third largest group) of the population of Kenya respectively. Because of their position of power, in a lot of sociopolitical issues these two ethnic groups (either jointly or individually) get pitted against the rest of the population. So, in that binary perspective (e.g., Kikuyus vs. "others"), the ethnic group in power was considered an ethnic minority by one annotator. The other annotator did not share this perspective and labeled these ethnic groups as indigenous groups. After a series of discussions with the annotators, a consensus was reached that the ethnic groups in power will be labeled neither as ethnic minorities nor as indigenous groups, but as a new target label: "large ethnic groups". This entailed that re-labeling of the extremity of these passages should take place.

E Model Details

Transformer models were finetuned for 3 epochs (5 minutes each), LSTMs for 5 and SVMs until convergence. A maximum length of 128 was used universally. For each baseline, three runs were made with results averaged. Standard deviations were minimal and were not reported for brevity.

The BERT-based models we used are:[6]

1. bert-base-multilingual-cased: https://huggingface.co/bert-base-multilingual-cased

	All		
	Der.	Exc.	Dan.
mBERT	84.9	55.1	50.4
	M	R	
mBERT	85.5	56.8	
	Target Group		
mBERT	91.4		

Table 13: Combined multilingual setting results.

2. bert-base-portuguese-cased: https://huggingface.co/neuralmind/bert-base-portuguese-cased

3. bert-base-german-cased: https://huggingface.co/bert-base-german-cased

4. hindi-bert: https://huggingface.co/monsoon-nlp/hindi-bert

5. bert-base-uncased-swahili: https://huggingface.co/flax-community/bert-base-uncased-swahili

F Combined Multilingual Setting

We perform an ablation study by combining all sets across countries and repeating our mBERT experiments in this new multilingual task (Table 13).

Even though the use of a "catch-all" model that is able to work on all languages sounds enticing, care should be taken to ensure that the model has sufficient understanding for each language and culture instead of making predictions based on dubious statistical cues (McCoy et al., 2019). This is a task out of scope for this work, but we are adding such a model to our baselines for completion.

G Test Set Results

In Tables 16, 17, 18, 19 and 20 we show results on the test set for tasks defined in §4.

[6] https://huggingface.co/models

	κ	α	ICC$(3,k)$	Targets	Ovr.	Der.	Exc.	Dan.	M	R
Overall	*0.23*	*0.24*	*0.41*	*0.69*	*63.0*	*78.4*	*40.2*	*18.8*	*78.4*	*46.3*
Brazil	0.08	0.12	0.19	0.62	85.9	91.3	12.7	5.8	91.3	6.7
Germany	0.35	0.35	0.52	0.79	68.2	73.0	61.6	0.0	73.0	61.7
India	0.11	0.04	0.19	0.81	39.6	72.2	30.2	5.3	72.2	39.7
Kenya	0.13	0.21	0.47	0.50	58.1	69.4	11.8	57.1	69.4	43.0

Table 14: Inter-annotator agreement table. In order, κ, α and ICC$(3,k)$ for extreme speech labels, target groups (κ), overall accuracy (%), derogatory/exclusionary/dangerous (%), M/R (%).

	Brazil		Germany		India		Kenya		*Total*	
	n	%	n	%	n	%	n	%	*n*	*%*
Politicians	1105	59.6	778	69.8	273	67.6	2098	93.9	*4254*	*75.9*
Legacy Media	663	35.8	106	9.5	75	18.6	54	2.4	*898*	*16.0*
The State	55	3.0	171	15.4	20	5.0	74	3.3	*320*	*5.7*
Civil Society Advocates	30	1.6	59	5.3	36	8.9	9	0.4	*134*	*2.4*

Table 15: Distribution of institutions of power as targets of derogatory extreme speech

	Brazil			Germany			India			Kenya		
	Der.	Exc.	Dan.	Der.	Exc.	Dan.	Der.	Exc.	Dan.	Der.	Exc.	Dan.
SVM	99.7	2.7	27.7	68.7	65.8	0.0	66.8	34.6	70.3	91.4	35.6	34.3
LSTM	98.7	0.8	0.0	78.2	55.9	0.0	54.5	62.6	0.0	66.8	68.2	0.0
langBERT	99.7	2.7	37.7	71.1	69.5	0.0	85.6	6.6	74.4	83.3	38.5	45.3
mBERT	99.5	0.0	34.8	58.2	74.0	0.0	93.1	4.1	73.6	86.2	47.1	55.2
XLM-R	100.0	0.0	0.0	65.6	76.2	0.0	96.3	0.0	49.6	90.6	35.3	24.4

Table 16: F1 for EXTREMITY, the three-way extreme speech classification task on the test set

	Brazil		Germany		India		Kenya	
	M	R	M	R	M	R	M	R
SVM	99.7	19.3	68.3	67.4	57.8	76.3	87.3	53.8
LSTM	97.6	24.8	78.6	52.0	64.7	80.3	82.4	56.7
langBERT	99.7	29.3	72.3	69.3	71.9	76.1	86.7	50.8
mBERT	100.0	0.0	54.2	75.9	80.0	50.6	86.5	61.4
XLM-R	100.0	0.0	100.0	0.0	0.0	100.0	86.5	63.2

Table 17: F1 for REMOVAL, the two-way extreme speech classification task on the test set

	Brazil	Germany	India	Kenya
langBERT	95.7	91.0	82.3	86.0
mBERT	95.2	90.0	91.7	89.3
XLM-R	95.2	89.9	90.1	87.2

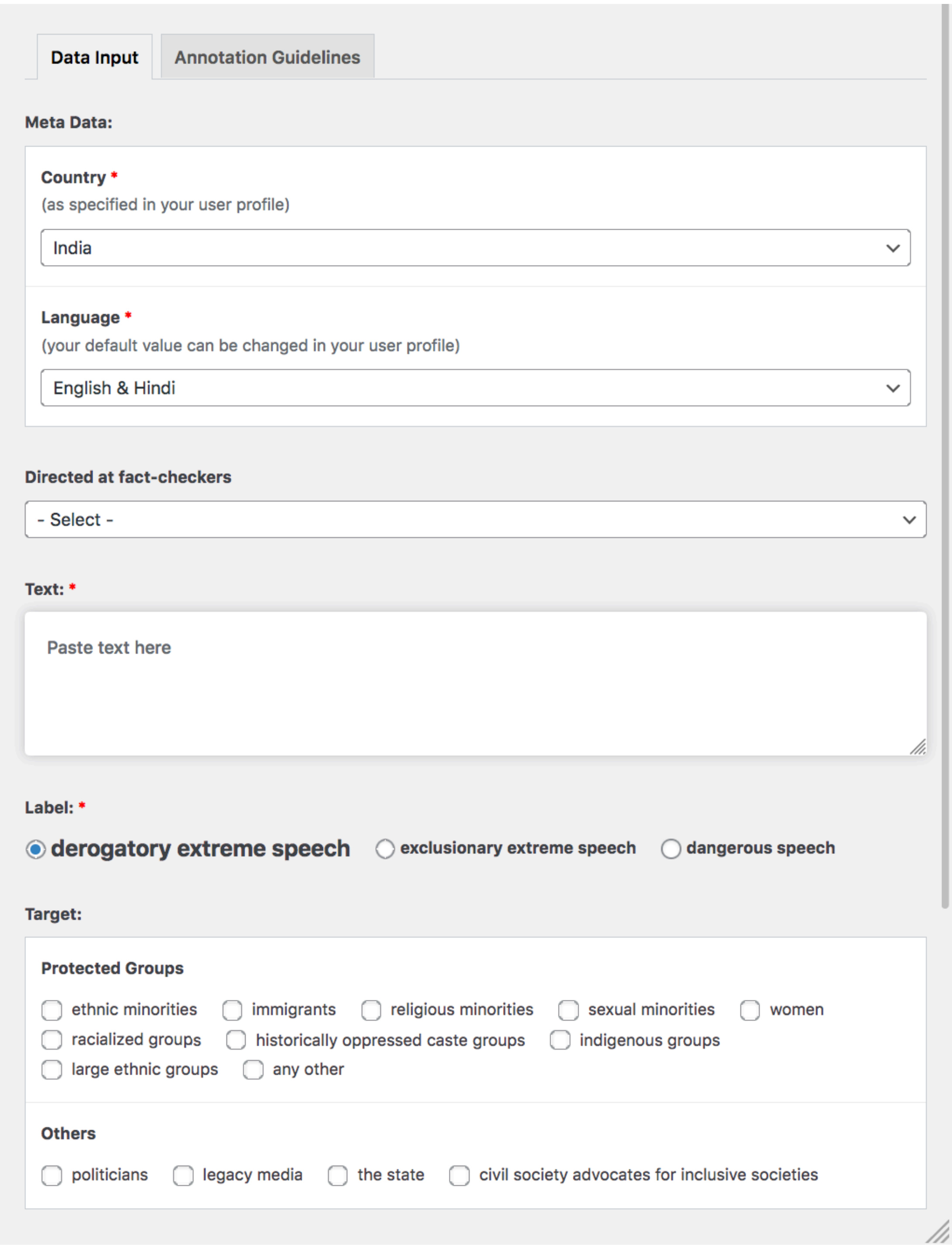
Table 18: LRAP (Label Ranking Average Precision) for TARGET, the target group classification task on the test set

		Brazil			Germany			India			Kenya		
		Der.	Exc.	Dan.	Der.	Exc.	Dan.	Der.	Exc.	Dan.	Der.	Exc.	Dan.
train	Brazil	99.5	0.0	34.8	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Germany	82.6	18.9	0.0	58.2	74.0	0.0	62.5	49.2	0.0	82.1	22.1	0.0
	India	63.9	5.4	31.9	56.2	37.2	0.0	93.1	4.1	73.6	69.7	34.6	9.0
	Kenya	95.2	0.0	2.9	82.7	7.2	0.0	79.4	8.2	32.0	90.6	35.3	24.4

Table 19: F1 for EXTREMITY in cross-country transfer (all languages) on the test set

		India$_{en}$			Kenya$_{en}$		
		Der.	Exc.	Dan.	Der.	Exc.	Dan.
train	India$_{en}$	60.0	69.0	50.0	62.1	45.4	0.0
	Kenya$_{en}$	83.3	4.0	18.8	84.3	62.1	55.1

Table 20: F1 for EXTREMITY for cross-country transfer in English on the test set



Meta Data:

Country *
(as specified in your user profile)

India

Language *
(your default value can be changed in your user profile)

English & Hindi

Directed at fact-checkers

- Select -

Text: *

Paste text here

Label: *

● derogatory extreme speech ○ exclusionary extreme speech ○ dangerous speech

Target:

Protected Groups

□ ethnic minorities □ immigrants □ religious minorities □ sexual minorities □ women
□ racialized groups □ historically oppressed caste groups □ indigenous groups
□ large ethnic groups □ any other

Others

□ politicians □ legacy media □ the state □ civil society advocates for inclusive societies

Figure 2: Interface presented to the annotators for data entry and labeling